DOCUMENT RESUME

ED 413 740 FL 024 761

AUTHOR	Zhiwei, Feng
TITLE	Language Technology and Language Resources in China.
PUB DATE	1995-00-00
NOTE	16p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Chinese; *Computational Linguistics; *Computer Software; Databases; Discourse Analysis; Foreign Countries; Grammar; *Ideography; Information Technology; *Language Processing; Language Research; Linguistic Theory; Machine Translation; Morphology (Languages); *Programming; Sentence Structure; Structural Analysis (Linguistics); Trend Analysis; Vocabulary; Written Language
IDENTIFIERS	*China

ABSTRACT

Trends and developments in computer applications in Chinese language research are described, focusing on these areas: input of Chinese characters and Chinese corpus; automatic segmentation of Chinese written text in corpus; development of a grammar knowledge base for Chinese words to be used as a resource for text segmentation and corpus annotation; automatic part-of-speech tagging for the corpus; automatic phrase bracketing and syntactic annotation for the corpus; creation of specialized terminology data banks; and machine translation systems. Specific projects, cooperative efforts, and resulting resources are noted. (MSE)

Language Technology and Language Resources in China

Feng Zhiwei

Institute of Applied Linguistics
State Language Commission of China
Chaoyangmen Nanxiaojie 51,
100010 Beijing, China
Fax: +86 106 513 8634

We are advancing into a new epoch – the information epoch. The remarkable feature of this information epoch is that the playing role of the computer in every aspect of society becomes more and more great. The natural language is the most important tool for communication of people, so it links an indissoluble bond with the information processing. In the information epoch, the computer with only a fourty year history gives a challenge to the Chinese language with six a thousand years history.

The Chinese language is the most important language of the Sino-Tibetan language family. Now nine hundred fourty million people in the world take Chinese language as their mother tongue. Not only Chinese people speak the Chinese language, some peoples in Singapore and Malaysia also speak it. The Chinese language is one of the working languages for the United Nations.

The Chinese character is the symbol set for recording the Chinese language. It is the largest symbol set of any writing system in the world. The Latin alphabet includes only 26 symbols, the Slavic alphabet 33 symblos, the Armenian alphabet 38 symbols, the Tamil alphabet 36 symbols, the Burmese alphabet 52 symbols, the Thai alphabet 44 symbols, the Lao alphabet 27 symbols, the Tibetan alphabet 35 symbols, the Korean alphabet 24 symbols, the Japanese Kana alphabet 48 symbols. However, there are too many symbols included in the Chinese characters. In the development procedure of Chinese characters, from ancient times to present times, the number of Chinese characters increased more and more. Following is the number of Chinese characters included in the dictionaries during different times of the Chinese history:

time	numbers
100 A.D. (Han Dynasty)	9353
543 A.D. (South Dynasty)	16917
1008 A.D. (Song Dynasty)	26194
1615 A.D. (Ming Dynasty)	33179
1716 A.D. (Qin Dynasty)	47043
1914	48000
1971	49888
1990	54678

With this big character set, how can the Chinese Natural language be processed by the computer? It is a great challenge to computational linguistics and corpus linguistics.

Forty years ago, in 1956, a Chinese scholar Ding Xilin suggested the creation of Šan, an electronic typewriter of Chinese characters. Prof. Qian Wenhao published a paper "Chinese characters and communication" in <Bulletin of Sciences>, in which he discussed the problem of encoding Chinese characters. In 1959, some Chinese scholars designed a machine translation system from Russian to Chinese (RC-59). It is the first connection between the Chinese language and the computer. In 1974, a large-scaled project, the 748 Project, was started: Chinese language processing became a new scientific subject in China.

In this paper, we shall introduce some results of Chinese language processing in China.

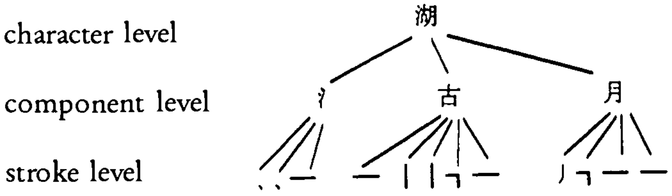
- Input of Chinese characters and Chinese corpus
- Automatic Segmentation of Chinese written text in corpus
- Grammar Knowledge Base for Chinese Words (GKBCW)
- Automatic POS (Part of Speech) tagging for Chinese corpus
- Automatic phrase bracketing and syntactic annotation for Chinese corpus
- Terminology Data-Banks
- Machine translation systems

1. Input of Chinese characters and Chinese corpus

The structure of a Chinese character can be divided in three levels:

character level
↑
component level
↑
stroke level

For example, 湖 (lake) can be divided as follows:



The Chinese character 湖 includes three components "氵古 月", and every component can include several strokes. The character is the higher level, the component is the middle level, the stroke is the lower level.

In the higher character level, the number of symbols is numerous, because different Chinese characters must be represented by different symbols. In the lower stroke level, the number of symbols is few, because we can sum up all strokes to several types. We can use only 5 kinds of strokes to represent all Chinese characters: horizontal stroke (–), vertical stroke (|), left-falling stroke (㇓), dot stroke (㇔), turning stroke (㇕).

In the middle component level, the number of symbols is not so great, however, it is also not so few. The statistical result shows: 11 834 different Chinese characters are composed only by 648 different components, 16 339 different Chinese characters are composed only by 673 different components. If we further sum up these different components to several component categories and encode all Chinese characters by fewer component categories, the component categories can be distributed in the keyboard of typewriter: we can have the possibility to input Chinese characters by the universal keyboard of the typewriter (QWERTY keyboard).

Of course, we can also encode all Chinese characters by 5 strokes, thus we can also input Chinese characters by the universal typewriter keyboard. In this case, we need only 5 keys on the keyboard.

How do you input Chinese characters into a computer? Now we have a simple method – the encoding method. We can encode Chinese characters by their strokes or components.

We can also encode the Chinese characters with the aid of the phonetic alphabet. In 1958, the Chinese government published "The Phonetic Scheme for Spelling Chinese Language". In 1982, this Scheme was accepted by ISO/TC46 as an International Standard – ISO 7098. We call this scheme PINYIN.

If we take PINYIN to represent Chinese characters, we can directly use the typewriter keyboard very well, because there is very good correspondence between PINYIN and the universal keyboard.

The problem for PINYIN input is homophone. The number of total syllables in the Chinese language is 408. These 408 syllables represent a large number of Chinese characters, so homophone is inevitable. However, many Chinese words are polysyllable words (bisyllable or trisyllable). If we take the word as an input unit, the number of homophones will be obviously decreased. Now a great many Chinese prefer the PINYIN input. I believe that PINYIN input will become the main tendency of keyboard input method for Chinese characters.

For information interchange, the State Bureau for Technique Supervision published the national standard "Character Set of Chinese Characters Encoding for Information Interchange - Basic Set" (GB 2312-80) in 1981. This standard includes two level of Chinese characters: the first level (commonly-used characters) includes 3755 characters; the second level (quasi-commonly-used characters) includes 3008 characters; the total number of characters included in GB 2312-80 is 6763. This is the basic set of Chinese characters.

In addition to the keyboard input method, there are the OCR input method and the speech-recognition input method.

OCR input method - The Optical Character Recognition (OCR) transforms the informations of Chinese characters in the paper into discrete electronic signals, then the computer recognizes these discrete electronic signals in order to recognize Chinese characters and to input Chinese characters into the computer. The Chinese OCR system can recognize 6763 Chinese characters in GB 2312-80; the recognition rate can reach 99.65%. The recognition speed can reach to 20 characters per second (in 386 PC).

Speech-recognition input method

- The computer directly transforms Chinese speech into Chinese text. The "SIDA-863A" system can recognize 398 basic Chinese syllables. The recognition rate can reach 93%. The response time is less than 0.1 seconds; input speed can reach up to 80 Chinese characters per minute. The syllable number of Chinese (420 syllables without ton, 1300 syllables) is fewer than English (4030 syllables) and Russian (2960 syllables), so Chinese speech recognition is relatively easy.

Since 1979, numerous Chinese corpora were created in China.

- corpus on Chinese contemparary literature (1979), 5.27 million Chinese characters, Wuhan University.
- Comprehensive Chinese corpus (1983), 20 million Chinese characters, Beijing Aviation & Spaceflight University.
- Corpus on Chinese language teaching materials for middle school (1983), 1068 million Chinese characters, Beijing Normal University.
- Comprehensive Chinese corpus (1983), 1.8 million Chinese characters, Beijing Language & Culture University.
- Corpus on Chinese Newspapers (1988), 2.5 million Chinese characters, Shanxi University:

  <People's Daily>: 1.5 million Chinese characters
  <Beijing Newspaper on Science>: 200 thousand Chinese characters
  <TV News> (CCTV): 500 thousand Chinese characters
  <Present Age> (magazine): 300 thousand Chinese characters.

- Chinese Corpus of Beijing University, begun in the early part of 1992, 5 million words.
- Bilingual (Chinese-English) corpus on computer science, begun in 1995 and is developing, State Language Commission, Applied Linguistics Institute.

From December 1991, a Chinese national Corpus project (sponsored by the State Language Commission) began to be put into effect. The purpose of this project is to build a large scale general corpus used for research of Chinese morphology, syntax, semantics, and pragmatics. This Corpus plans to collect 70 million Chinese characters.

The selection of this Corpus has the following restrictions:

- Diachronic restriction: to select the full range of materials from 1919 until now, and give priority to the material after 1977.
- Cultural restriction: mainly to select material that can be understood by the persons who have received formal education to the graduation of secondary school.
- Usage restriction: mainly to select the commonly-used material, and give priority to the material of social sciences and humane studies.

At present, this project progresses smoothly. By the end of 1995, the completed corpus will reach 20 million Chinese characters.

2. Automatic segmentation of Chinese writting text in corpus

A Chinese sentence is a sequence of Chinese characters, there are no obvious delimiting markers (such as spaces in Europaen languages) between Chinese words except for some punctuation marks. Because of this, word segmentation is essential in Chinese language processing.

In order to identify the embedded words in Chinese text, we match the input Chinese character string with the lexical entries in a large Chinese dictionary.

There are many matching methods:

- Maximum Matching method (MM method): taking 6-8 Chinese character string as maximum string, to match the maximum string against the lexical entries in dictionary, if failed, to cut one Chinese character and to match further until a corresponding word in the dictionary is found. The segmentation direction is from right to left.
- Reverse maximum matching method (RMM method): segmentation direction is from left to right. The experiment shows: RMM method is better than MM method.

- Bidirection Matching method (BM method): Comparing the segmentation resulted of MM and RMM, then to decide a correct segmentation.
- Optimum Matching method (OM method): In dictionary, the entries order was arranged by their frequency in the Chinese text, form higher frequency word to lower frequency word.
- Association-Backtracking method (AB method): By means of an association mechanism and backtracking mechanism to do the matching.

However, some ambiguous segmentation strings (ASSs) and unregistered words (the words that are not registered in the dictionary, URWs) in the text shall lower the accuracy of segmentation.

There are two types of ASSs:
- overlapping string: e.g., 太平淡 (very prosaic, rather flat): 太平 (peace and tranquility) + 平淡 (prosaic), 平 becomes overlapping segment.
- combinative string: 马上 e.g., (at once): 马 (horse) + 上 (upper)

= on the horse.

URWs are mainly proper nouns: personal name and place name. e.g, 冯志伟 (Feng Zhiwei), 蒂豪尼 (Tihany). These are not included in the dictionary.

To resolve these problems, various knowledge might have to be consulted. Knowledge on the part of speech (POS) will be helpful. If we blend the POS annotation and automatic segmentation, the accuracy of segmentaion will be increased remarkably.

For Chinese word segmentation, a natioanl standard was promulgated in 1992 – 'Contemporary Chinese language word segmentation standard used for information processing' (GB 13715). This national standard proposes the principle for determination of Chinese words. It is the basic principle for automatic word segmentation.

3. Grammar Knowledge Base for Chinese Words (GKBCW):

Now a GKBCW is developing in Peking University. This GKBCW can be used as a large electronic dictionary; it is a support to the automatic segmentation and Chinese Corpus annotation.

In GKBWC, each word category has several features, e.g., the verb category has about forty features, the noun category has twenty-five features. These features describe the grammatical functions and distribution of every Chinese word in the text. GBKCW includes about 50 000 word entries which composed a general base and 27 sub-bases.

The general base describes the common features for all words, such as

- pronunciation of word: e.g., 太平 [tai4ping2], 马上 [ma3shang4].
- part of speech of word: e.g., 编辑 can be verb (to edit) or noun (editor), 制服 can be verb (to bring under control) or noun (uniform).
- type of ASSs: e.g., 太平 can lead to overlapping string, 马上 can lead to combinative string.
- usage features: e.g., frequency, subject domain, style, rhetoric feature.
- special information: e.g., radicals of Chinese character by which characters are arranged in traditional Chinese dictionaries. 纟 in Chinese character 编 is a radical in the traditional dictionary.

The 27 sub-bases are for basic parts of speech, idiom, some Chinese characters which can not be used as morpheme, and Chinese punctuation markers, etc.

The general base is linked with 27 sub-bases. All the information in the general base can be transfer to each sub-bases.

4. Automatic POS tagging for the Chinese Corpus:

There are two kinds of POS tagging approach:

- Statistics-based approach
- Rule-based approach

The processing procedure for statistics-based approach of POS tagging can be divided into the following steps:

- Manually analyse some texts selected from corpus (training set), annotate the training set, and extract the statistical data from the analysed training set (it is represented by 2-tuple grammar);
- construct a statistical model according to the results from the statistical data of training set;
- automatically annotate new texts based on the statistical model.

All tagging information is coming from the electronic dictionary where is recorded the information about the POS for every word.

The automatic POS tagging system of Qinhua Unversity in Beijing takes the statistics-based approach; the accuracy of POS tagging reached 96.8%; the annotating speed is 175 occurances per second.

For the rule-based approach, a serious problem is the POS ambiguity. In Chinese language, POS ambiguity mainly concentrates on the frequently-used words: verb, noun adjective, etc.

verb-noun ambiguity:	37.60%
verb-adjective ambiguity:	24.30%
noun-adjective ambiguity:	10.40%
adjective-adverb ambiguity:	4.55%
verb-preposition ambiguity:	4.04%
verb-adverb ambiguity:	2.27%
noun-verb-adjective ambiguity:	2.27%
noun-adverb ambiguity:	2.02%
other ambiguity:	12.55%

The disambiguation must be based on linguistic rules (grammar, semantics, context, etc), so that the information included in GBKCW will be very helpful.

In fact, the statistics-based approach is an empirical approach, and the rule-based approach is a rational approach. We can combine both statistics-based approach and rule-based approach into one, and integrate different types of approaches in POS tagging. By this approach, experimental results of Peking Unversity are:

- segmentation accuracy: 97.68% (close test)
- POS tagging accuracy: 96.06% (close test), 95.72% (open test).

5. Automatic phrase bracketing and syntactic annotation for the Chinese Corpus

After word segmentation and POS tagging for Chinese corpus, we must manually proofread the results, and when we can confirm that the results are good, then we can start the automatic phrase bracketing and syntactic annotation.

The procedure is as follows:

- To predict the boundary locations of a phrase according to the information about words, their POS and other syntactic features, and determine which word is the left boundary of a phrase, which is the right boundary of a phrase, which word is the middle part of a phrase.

We can bracket a phrase as following:

[w w ... w w]

[w is an open bracket, w] is a closed bracket.

- To match the open brackets and its corresponding closed brackets based on the context information.

- To resolute the ambiguity of the phrase according to the disambiguation rules and statistical information.
- To generate the constituent structure tree for a sentence.

Recently a Chinese Corpus Multilevel Processing system (CCMP) is being developed at Peking University. This CCMP system includes two sub-systems and supplementary tools:

- Word segmentation and POS tagging sub-system
- Phrase bracketing and syntactically annotating sub-system
- supplementary tools such as query tools, sample tools, statistical tools, and corpus management interface.

The experimental results:

- the percentage of crossing brackets: 13.98%
- the percentage of error phrase tags: 8.65%

That means there are many problems to resolve for the annotation of Chinese corpus.

6. Terminology Data-banks

The terminology is crystallization of scientific knowledge in language, it is an important language resource.

In 1990, the sub-committee of computer-aided in terminology of China was set up. This sub-committee is attached to the State Language Commission (SLC) of China.

A series of national standards for a terminology data-bank are promulgated:

- General principles and methods for establishing terminology data bank, 1992, GB/T 13725-92.
- Magnetic tape exchange format for terminological/lexicographical records, 1992, GB/T 13725-92.
- Guideline for the development of terminology data banks, 1993.
- Guideline for the documentation for developing terminology data bank, 1993.
- Guideline for the evaluation of terminology data banks, 1994.

Many terminology data-banks are created:

- GLOT-C: data processing termonology, Chinese-English, 1988, the Academia Sinica collaborated with FhG of Germany.
- TAL: applied linguistics termonology, Chinese-English, 10 000 terms, the State Language Commission, 1990.

- COL: computational linguistics terminology, Chinese-English, 10 000 terms, the State Language Commission collaborated with Trier University of Germany, 1993. (in press by Langenscheidt Verlag).
- Terminology data-bank on computational linguistics: Chinese-English-German-Japanese, 12000 terms, Peking Unversity, 1994.
- Terminology data-bank on machinary: 250 000 terms, Chinese-English-French-German-Russian-Japanese, Institute for Scientific and Technical Information, the Ministry of Machinary, started since 1989, in development.
- Thesaurus bank on agriculture: Chinese-English, 25 000 terms, Chinese Academy for Agriculture, 1991.
- Thesaurus bank on chemical industry: Chinese-English, 25 000 terms, China Information Center of Chemical Industry, 1989. There are two versions: published version and machine-readable (floppy discs) version. All terms can be transmitted through the network to provide information retrieval service.
- Encyclopedia terminology data bank: Chinese-English, there are definition and explanation for every term, 180 000 terms, China Encyclopadia Press, 1995.
- Terminology data bank for standardization: Chinese-English, it includes several sub-databanks:

. comprehensive terminology data bank (TDB)
. comprehensive bibliographic data bank (BDB)
. comprehensive factual data base (FDB)
. filing system administration in the form of a comprehensive data-bank (FSA)
. multiple documentation language system (MDL)
. office automation system (OAS)
. full text system (FTS)
. data base for graphical and other non-linguistic data (GDB)

China Standardization and Information-Classification-and-Coding Institute (CSICCI) collaborated with Oesterreichisches Normungsinstitut, in developing.

- Project of Comprehensive scientific terminology data bank: Chinese-English, 50000 terms, Institute of Scientific and Technical Information of China (ISTIC), the project started in 1995, shall fulfill in 1998.

7. Machine Translation Systems

The study of machine translation (MT) in China started over forty years ago. The development of MT in China can be described in four periods:

- the early experimental period (1956–1966)
- the stagnant period (1966–1975)
- the recovery period (1975–1987)
- the blossom period (since 1987)

In 1956, the MT research has been included in the National Plan for Developing the Science and Technology as a project named "Machine Translation – establishing of the translation rules of natural language and mathematical theory for natural languages". This project can be divided into two parts: one is the establishment of the translation rules of natural language – "machine translation", another is the study of mathematical theory for natural language – "mathematical linguistics" (the theoretical foundation for machine translation). Several research groups were founded in Beijing (Academia Sinica, Beijing Institute of Foreign Language), in Guangzhou (South China Polytechnical Institute), and in Harbin (Harbin Polytechnical University). In 1959, a Russian-Chinese MT experiment (RC-59) was successfully fulfiled on a general-purpose computer 104. With a vocabulary of 2030 Russian words, an algorithm of 29 flowcharts, this RC-59 experiment encouraged the belief that MT from foreign language into Chinese is possible. At the same time, some scientists began to study the MT from English to Chinese. In 1960, an English-Chinese MT algorithm was composed. A monographical brochura "Preliminary of Machine Translation" was published in 1965. However, from 1966 until 1975 the MT in China completely stagnated.

Since 1975, after a long sleep of 10 years, the MT in China restarted and came to the recovery period.

In November 1975, a MT joint-research group was established. This MT group consists of the Institute of Scientific and Technical Information of China (ISTIC), the Linguistics Institute of the Chinese Academy for Social Sciences (CASS), the Computational Technique Institute of Academia Sinica etc. This group carried out a MT experiment from English to Chinese on the TK-70 computer and T-4100 information processing device of Chinese characters. The raw materials contain 9200 English titles of scientific and technical papers. As the results of this MT experiment, a MT system TITLE-1 was set up in 1986.

At the same time, MT study was carried out in Helongjiang University (Harbin), in the Mars Institute (Beijing), in the Telecommunication

University (Beijing), in South China Polytechnical University (Guangzhou), in Central China Polytechnical University (Wuhan), in the Institute of Scientific and Technical Information of Shanghai (Shanghai), and in Inner Mongolia Unversity (Huhehot).

In this recovery period, interactive approaches and diverse strategies were developped, some multilingual systems appeared, and the application of AI technique in MT began to be considered. Mathematical linguistics were also studied in the universities or the institute. A monograph "Mathematical Linguistics" was published in 1985.

For the sake of investigation of linguistical phenomena, an English corpus was created in Jiaotong University (Shanghai), and numerous Chinese corpora were created in Wuhan University, Peking University, Qinhua University, Shanxi University.

In the recovery period, most of the MT system was experimental:

- TITLE-1 system: English-Chinese, ISTIC, 1976-1986.
- ECMT-1 sytem: English-Chinese, Linguistics Institute, CASS, 1978.
- JFY system: English-Chinese, Linguistics Institute, CASS, 1976-1984.
- INSPEC sysetm: English-Chinese, Telecommunication University, 1985.
- HT-83 system: English-Chinese, Helongjiang Unversity, 1983.
- RI-84 system: English-Chinese, Helongjiang University, 1984.
- GCAT system: German-Chinese, Applied Linguistics Institute, LSC, 1985.
- FCAT system: French-Chinese, Applied Linguistics Institute, LSC, 1985.
- FAJRA system: Chinese-French/English/Japanese/Rassian/German, Applied Linguistics Institute, LSC, 1981.

The TITLE-1 system possessed a large-scale electronic dictionoary including a basic dictionary (20 000 entries) and an idiomatic dictionary (67 000 entries). This system can translate the English titles of scientific papers in the field of metalurgy to Chinese, the average translation speed is 80 titles/hour.

Since 1986, the MT of China came to the blossom period. The symbol of this blossom period is the KEYI-1 English-Chinese system of the Mars Institute (Beijing). In March 1987, KEYI-1 system passed the academic appraisal by experts. Its translation ability is as the ability of graduated students of the English department in China, its translation speed is 3000 words/hour, and the result of translation is readable. In the process of machine translation, the user can input their special words to KEYI-1 in order to adapt to their special demands.

KEYI-1 system quickly became an operational system and was commercialized; China National Software & Technology Service Co. (CS&S) bought the copyright of the system, and KEYI-1 system was renamed as TRANS-STAR system. CS&S put it on the market and gained the profit.

Now TRANS-STAR system has been improved. It is now much better than KEYI-1. The translation speed is raised to 15 000 words/hour for 286 PC, 30 000 words/hour for 386 PC. The basis dictionary includes 40 000 entries; the system has 10 specialized technical dictionaries including 350 000 entries. The subject fields involved computer, economics, telecommunication, ceramics, thermal power industry, printing machinary, automobile and tractor industry, petroleum prospecting, geology, and chemical industry.

In the blossom period, another three operational systems are also very successful:

- GAOLI MT system (English-Chinese): It is jointly developed by Beijing GAOLI Computer Co. Lid. & Linguistics Institute of CASS.
- basic lexical base: 60 000 entries in which the usage of every word is described.
- linguistic rules: more than 800 rules used for syntactic analysis of English and generation of Chinese.
- background knowledge base: more than 150 entries used for semantic analysis and generation
- translation accuracy: 80%
- readability of translated text: 80%-90%
- translation speed: 12000 words/hour for 386 PC.
- 863-IMT/EC system (English-Chinese): It is developed by the Computer Technology Institute, Academia Sinica. This system was commercialized and earned very good economic benefits.

. basic English lexical base: 35 000 entries
. basic Chinese lexical base: 25 000 entries
. linguistical rules: 1500 rules
. translation accurancy: 80%

- SINO-TRANS system (Chinese-English): It is developed by CS&S at 1993.

. basic dictionary: 40 000 entries
. two specialized technical dictionaries: navel ships and boats (9312 entries), rocket gun (33 773 entries)
. linguistic rules: 1000 rules
. translation speed: 20 000 Chinese characters/hour.

Since 1989, the corpus approach (e.g., statistical approach, example-based approach) is introduced to machine translation, all the research work of machine translation are based on the processing of large-scale authentic corpus. The combination of machine translation with the corpus approach will promote the development of Chinese language technology. Corpus linguistics play more and more a role in Chinese language technology. The prospect of Chinese language technology will be more and more brilliant.

References

Feng Zhiwei. 1982. "Memoire pour une tentative de traduction multilingue du chinois en francais, anglais, japonais, russe et allemand" Proceedings for COLING'82, Prague.

Feng Zhiwei. 1983. "Multi-label and multi-branch tree for automatic analysis of Chinese sentences". Proceedings for 1983' International conference on Chinese information processing, Beijing.

Feng Zhiwei. 1984. "Automatic generation and analysis of Chinese language in machine translation". Proceedings of SEARCC'84, Hongkong.

Feng Zhiwei. 1987. "Linguistic information included in Chinese sentences". Proceedings of TKE'87, Trier.

Feng Zhiwei. 1989. "Some special problems of machine translation in China". Proceedings for Chinese Computing Conference'89, Singapore.

Feng Zhiwei. 1990. "Complex features in description of Chinese language". Proceedings for COLING'90, Helsinki.

Feng Zhiwei. 1990. "Automatic analysis of Chinese – MMT model". Proceedings of 1990 International Conference on Computer Processing of Chinese and Oriental language, Beijing.

Feng Zhiwei. 1991. "On potential ambiguity in Chinesec terminology". Proceedings of TSTT'91, Beijing.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: TELRI - Proceedings of the First European Seminar:"Language Resources for Language Technology", Tihany, Hungary, Sept. 15 and 16, 1995

Author(s): Heike Rettig (Ed.)

Corporate Source:

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	The sample sticker shown below will be affixed to all **Level 1** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 1**	The sample sticker shown below will be affixed to all **Level 2** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 2**	□ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title:
Norbert Volz, M.A.
TELRI Project Manager

Organization/Address:
Institut für deutsche Sprache
R 5, 6-13 - 68161 Mannheim
Postfach 101621 - 68016 Mannheim

Telephone: +49 621 1581-437

FAX: +49 621 1581-415

E-Mail Address: volz(at)ids-mannheim.de

Date: 28/11/97

(over)